Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/2/8/e1600451/DC1

Supplementary Materials for

Humans display a reduced set of consistent behavioral phenotypes in dyadic games

Julia Poncela-Casasnovas, Mario Gutiérrez-Roig, Carlos Gracia-Lázaro, Julian Vicens, Jesús Gómez-Gardeñes, Josep Perelló, Yamir Moreno, Jordi Duch, Angel Sánchez

Published 5 August 2016, *Sci. Adv.* **2**, e1600451 (2016)
DOI: 10.1126/sciadv.1600451

This PDF file includes:

- fig. S10. Average fraction of cooperative actions (and SEM) among the population as a function of the round number overall (left) and separating the actions by game (right).
- fig. S11. Distribution of fraction of rational actions among the 541 subjects of our experiment, when considering only their actions in HG or PD, or both.
- fig. S12. Fraction of rational actions as a function of the round number for the 541 subjects, defined by their actions in the PD game and HG together (top) and independently (bottom).
- fig. S13. Values of risk aversion averaged over the subjects in each phenotype.
- fig. S14. Average response times (and SEM) as a function of the round number for all the participants in the experiment and separating the actions into cooperation or defection.
- fig. S15. Distributions of response times for all the participants in the experiment and separating the actions into cooperation (top) and defection (bottom).
- fig. S16. Testing the robustness of the results from the *K*-means algorithm.
- fig. S17. Davies-Bouldin index as a function of the number of clusters in the partition of our data (dashed black) compared to the equivalent results for different leave-*p*-out analyses.
- fig. S18. Average value for the normalized mutual information score, when doing pairwise comparisons of the clustering schemes from 2000 independent runs of the *K*-means algorithm both on the actual data and on the randomized version of the data.
- fig. S19. Age distribution for the different phenotypes compared to the distribution of the whole population (black).
- fig. S20. Difference between the experimental (second row) and numerical (or inferred; first row) behavioral heatmaps for each one of the phenotypes found by the *K*-means clustering algorithm, in units of SD.
- fig. S21. Average level of cooperation over all game actions and for different values of T (in different colors).
- fig. S22. Average level of cooperation as a function of (T,S) for both hypothesis and experiment.

S.1. Technical implementation of the experiment

To conduct the experiment and collect the data we implemented a local network architecture (see fig. S1) which consisted of 25 mobile devices (tablets), a router, and a laptop running a web server and a database server. The system was designed to allow playing synchronized sessions, to collect and store user data safely, and to control in real time the experiment while the users were playing against each other.

The game was accessible through a web application specifically designed for tablets. All the interactions that users made through the game interface were immediatelly sent to the server through a client API -no data was stored in the tablets-. The server also provided a server API to control and monitor the status of each experiment session.

The software of the experiment was developed using Django framework and Javascript. Both APIs were implemented using RESTful services and JSON objects for the exchange of data between server and clients, which was stored in a MySQL database.

S.2. Running the experiment

The experiment was carried out during the game festival (Festival del Joc) DAU Barcelona `http://lameva.barcelona.cat/daubarcelona`, in December 2014, over a period of two days. We collected data from 541 subjects in total, who were recruited by our team among the game fair attendees. Due to space limitations, the experiment took place in multiple sessions over those two days, in groups of $15 - 25$ people. The average age among our 541 subjects was 31.3 (SD=14.3) (see fig. S2 for the age distribution of the population), with 64.5% males and 35.5% females.

Each person was given a tablet to play the game using the tablet's browser. Before the actual experiment started, the subjects were shown a tutorial in their tablets, to learn (i) the basic rules

of how to play the game, (ii) an explanation about the meaning of the payoff matrix and their possible choices, and (iii) a couple of examples of game rounds equivalent to the ones they would face during the actual game. Also, some of our team members were walking around the room answering questions from the subjects during the tutorial period (but not during the actual game). Nonetheless, we did not instruct them to play in any particular way nor with any one particular goal in mind. In fig. S3 we show the tutorial screens. After a player had read the tutorial, she pressed a button to indicate the system that she was ready to start playing. Once everyone was ready, the game administrator started the game.

Each game session was carried out for a random number of rounds, between 13 and 18. The players did not know the total number of rounds they were going to play. For each round, subjects were randomly assigned different opponents, and nobody knew who they were playing against. In each round of the game, the players had 40 seconds to make their action choice. If they did not choose anything, a random choice was generated by the system (and saved in our database, properly labeled to be discarded in the analysis). After a player had made a decision in a particular round, she had to wait until all other players were done too, before obtaining the outcome of the round and proceeding to the next game round (fig. S3j). Finally, in order to encourage the experimental subjects' decisions with real material (economic) consequences, they were informed that they would receive lottery tickets proportionally to the payoff they accumulated during the rounds of dyadic games they played. The four prizes in the corresponding lottery were coupons redeemable at participating neighboring stores, worth 50 euros each.

S.3. Translated transcript of the tutorial and feedback screen after each round

Before the experiment started, and for each group of subjects, we showed them a tutorial in the same tablets used to play the game. The format and presentation of the game examples used

in the tutorial were identical to those of the real experiment. We present next the translation into English of the text from every screen of the tutorial (the original was made available to the participants in Castilian/Spanish and Catalan.

Tutorial Screen #1. See fig. S3(a). *Welcome to Dr. Brain. The game, designed to study how we make decisions, is made of several rounds with different opponents located in the DAU. During the experiment we don't expect you to behave in any particular way: there are no wrong nor incorrect answers. You will simply have a limited time to make your decisions. In these next screens we will teach you how to play Dr. Brain. Use the side arrow keys to move within the tutorial, and when you are done you will be able to start the rounds. This game has been thought by scientists from the Universitat of Barcelona (UB), Universitat Rovira i Virgili (URV), Instituto de Biocomputación and Sistemas Complejos (BIFI)-Universidad de Zaragoza (UZ) and Universidad Carlos III in Madrid (UC3M). It is an experiment to study and understand how we humans make decisions.*

Tutorial Screen #2. See fig. S3(b). *The rules of Dr. Brain. It is important that you don't talk to other players during the experiment. Keep focused! The decisions made during the experiment and the accumulated points will determine your chances of wining prizes: the more points, the more tickets you will get for the raffle. If you leave the game while it is in progress, you won't be able to come back in!*

Tutorial Screen #3. See fig. S3(c). *This is the screen you will see when the rounds of the game start. In each one of them, we will assign you a random partner to play.*

Tutorial Screen #4. See fig. S3(d). *Each round has a table that represents your opponent's possible actions as well as yours. Your opponent and you will follow the same rules in the round. In this way, depending on what each one of you choose, you will win more or less. The rows represent your choice, the columns represent your opponent's. For each choice, it is listed how much you will win, and how much your opponent will.*

Tutorial Screen #5. See fig. S3(e). *Pay attention, the tables may change from round to round, and the rules may be different. You may win more or less points, o what seemed more interesting may be different now.*

Tutorial Screen #6. See fig. S3(f). *To play you must choose one of the two options, represented by a color. Your opponent plays following the same rules as you, described in the table, but you won't know his choice until after the end of the round.*

Tutorial Screen #7. See fig. S3(g). *Every round of the game lasts 40 seconds, you have to choose one of the two actions during that time. If you don't choose anything, the computer will do it for you randomly and you will move on to play the next round. Don't worry, 40 seconds is plenty of time!*

Tutorial Screen #8. See fig. S3(h). *Example: If you pick RED and your opponent picks GREEN. You (red) win 8 and your opponent (green) wins 6.*

Tutorial Screen #9. See fig. S3(i). *Example: If you pick PURPLE and your opponent picks YELLOW. You (purple) win 11 and your opponent (yellow) wins 0. If your adversary chooses... If you choose... You win... He wins... What do you choose?*

Feedback Screen after a typical round of the game. See fig. S3(j). *Almost there, thanks for your patience! You and your opponent have both chosen YELLOW. You and your opponent have earn 5 each. Next game starts in... (countdown)*

S.4. Other experimental results

S.4.1. Fraction of cooperation by age and gender

We did not find any significant differences in the fraction of cooperative actions in the whole (T, S)-plane by age when separating young players (≤ 15 years old) from adults (> 16 years old) (see fig. S4) nor between males and females (see fig. S5).

S.4.2. Fraction of cooperation by game round

We did not observe large differences in the fraction of cooperative actions in the whole (T, S)-plane when separating by game round, with the exception of the first few rounds of the session (see fig. S6 for heatmaps of cooperation and fig. S7 of heatmaps of relative differences in cooperation).

S.4.3. Number of actions per (T, S)-plane point and Standard Error of the mean fraction of cooperation

The total number of actions generated by our 541 subjects was $8,366$. The (T, S)-plane was discretized into a 11×11 lattice, and the (T, S) point for any given pair of opponents and for any given round was randomly generated in such a way that subjects had uniform probability to be assigned to any point in the (T, S)-plane. Thus, the average number of actions per (T, S) point is 69. In fig. S8 we show the total number of actions per point in the (T, S)-plane for all subjects.

On the other hand, in fig. S9, we show the Standard Error of the mean fraction of cooperative actions for all the actions and all the players in the experiment, for the whole (T,S)-plane. We observe that the values for the Standard Error of the mean are uniformly distributed across the entire (T,S)-plane, except for the upper-left triangle of the HG, where the error is clearly lower than in the rest of the regions. This seems to indicate that at a population level, most people chose the same action at least in that particular region.

S.4.4. Time evolution of the fraction of cooperation

Our experiment was designed to avoid learning or memory effects as much as possible, making each subject play knowingly in different game conditions and against different anonymous opponents in every round. In the left panel of fig. S10, we show the average fraction of cooperative

actions as a function of the round number over the whole population, and we observe how there is only a very small decline in cooperation as the round number increases, specially during the first two or three rounds. Also, note that the dispersion of the values is larger in the last few rounds, since every subject play a random total number of rounds between 13 and 18 rounds. Similarly, we show in the right panel of fig. S10 the average fraction of cooperative actions as a function of the round number, separating the actions into the different games. In this case we do observe a small decline of cooperation in the case of the Prisoner's Dilemma (PD) and the otherft (SG), and a small increase in cooperation in the Harmony (HG), while the fraction of cooperative actions doesn't show any particular trend for the Stag Hunt (SH).

S.4.5. Rationality and Risk aversion

We measure the level of rationality (only under the assumption of self-interest) among our subjects using only their actions in the Harmony and/or Prisoner's Dilemma games. According to Game Theory, the rational action in the Harmony game is to cooperate, while in the Prisoner's Dilemma it is to defect.

In fig. S11 we show the distributions of the fraction of rational actions chosen by the subjects in the Harmony game (HG), in the Prisoner's Dilemma (PD), and in both games combined, along with the corresponding mean values among the population (vertical purple lines). We observe that an important subset of individuals presents a fraction of rational actions near 1.0 (around 50% of subjects when calculated with either game independently, and around 30% when calculated with both games combined). However, there are also some others that act irrationally (around 5% or 10% as calculated with either game). Note that the average value of rationality of the whole population when both games are considered in the statistics, is around 75% (see purple vertical lines in fig. S11).

Moreover, we checked the time evolution of the fraction of rational actions in the population,

as defined by their actions in the Harmony (HG) and Prisoner's Dilemma (PD) games together, and independently (fig. S12), and we do not observe any significant increase or decrease of rationality as a function of the round number in any case.

Regarding the definition of risk-aversion, we choose to define it as the number of cooperative actions in the SG together with the number of defective actions in the SH (over the total sum of actions in both quadrants for a given player). The rationale behind such a combined measure of risk aversion is the avoidance of the bias of pure cooperativeness: were we to measure risk aversion only in the SH (instead of combining both SH and SG), for a group that defects a lot everywhere in the (T, S)-plane, it would appear as if they are more risk averse than they really are, while a mostly cooperative group would appear as less risk averse than they really are. A similar reasoning would apply to only using the SG quadrant for the measure, and therefore we have looked at the actions in both the coordination and anti-coordination games together.

In fig. S13 we represent the average values of risk-aversion according to this definition, for each one of the phenotypes, and the population as a whole. While Envious, Trustful, and Un players exhibit intermediate levels of risk aversion (0.52, 0.52 and 0.54, respectively), Pessimists exhibit a significantly higher value (0.73), consistent with their fear of facing the worst possible outcome and their choice of the best worst-case scenario. In contrast, the Optimist phenotype shows a very low risk aversion (0.32), in agreement with the fact that they aim to obtaining the maximum possible payoff, risking the possibility that their counterpart do not work with them towards that goal.

S.4.6. Response times

We have also examined the response times of the individuals in our experiment, separating the data by cooperation/defection actions, and as a function of the round number. Figure S14 shows that the average response time is around 15 seconds. We did not find any dependence with the

round number nor with the type of action. Finally, fig. S15 displays the distributions of response times for all individuals, for each of the two possible actions.

S.4.7. Clustering Analysis

We hypothesized that there are distinct, well-defined types of individuals (or phenotypes) in our dataset, that can be told apart by using an unsupervised clustering algorithm. Hence, we run a K-means clustering algorithm on our data (using the Scikit-learn Python package) to analyze its clustering structure. We represent each participant in the dataset by a four-dimensional vector, corresponding to her average fraction of cooperative actions in each one of the four dyadic games (Prisoner's Dilemma, Stag Hunt, Snowdrift and Harmony).

The K-means unsupervised clustering algorithm groups the data into a user-defined number of clusters, by both minimizing the dispersion within each cluster and maximizing the distance between the centroids of each pair of clusters. For a given number of clusters, $k = 2, 3, 4, ..., 20$, we run the algorithm 200 times on our data (with different seeds for the algorithm in every run), and obtain the average value of the BD-index (see subsection below for formal definition), which is a measure of how optimal is that K-scheme. This way we can pick which one is the best cluster scheme. In fig. S16 we show the average value and the Standard Deviation (SD) of the DB-index, as a function of the number of clusters in the partition. This representation will have a minimum around the optimum number of clusters for a given dataset. Conversely, it would be monotonically decreasing if the data set lacks any significant cluster structure.

We found that there is an optimum around a scheme with 5 or 6 clusters (black line in fig. S16). However, due to the fact that the SD is considerably smaller for 5 than for 6 (which indicates that the partition schemes found in different realizations of the algorithm for $k = 5$ are much more similar to each other in terms of their corresponding DB-index, than in the case of $k = 6$), we pick $k = 5$ as our optimum clustering partition. Note also that the SD is very

large for any partition with 6 or more clusters, which also points to the lack of robustness of those partition schemes.

It is also important to mention that this clustering approach does not allow us to compare our results against the 'ground truth', since that is unknown to us. We can only test for its robustness, and we do this in multiple ways. We present the results from the same algorithm, also run 200 times, but this time on a randomized version of our data. This data randomization is done as follows: we take the $8,366$ actions of the 541 subjects and create an 'action pool' with them. From this pool of data we draw (with replacement) to obtain the new, randomized sets of actions for each person, in such a way that we preserve the number of times each subject has played and the particular (T, S) points she played in, but now her actions are randomized. With this randomization procedure we preserve the average fraction of cooperative actions in the population, but destroy any possible correlations among the actions of any given subject. Note in fig. S16 that with the randomized version of the data (green line), there is no local minimum for the DB index, and the best partition would be to have as many clusters as possible, which is an indication of the lack of internal structure of the randomized data.

On the other hand, and recalling that the cooperation patterns in the heatmap for all users seems to be a little less clear during the first few rounds (while the subjects seem to be picking up the mechanics of the experiment), than during the rest of the experiment (see fig. S6), we also test the clustering structure of our data when removing the first couple of rounds for every subject. In this case, we observe that the cluster structure is even clearer, with an even more significant minimum at $k = 5$ clusters, as indicated by the DB-index (fig. S16, red line).

On the other hand, we also wanted to test the robustness of our clustering analysis against data perturbations, specifically by running it on just a subset of the original data. In order to do so, we run the algorithm 200 times again, but in each realization we exclude a given number of players and all their actions, randomly chosen (that is to say, we perform a leave-p-out analysis,

for different values of p). We do this for a scheme with $k = 2, 3, 4, ..., 20$ clusters, and leaving out $p = 100, 300, 400$, and 450 subjects (out of the total 541), and calculate again the average DB index for them. In fig. S17 we show the results from the leave-p-out procedure as they compare to the original data (the black dashed line in fig. S17). We observe that the results of the K-means analysis in our data are very robust when randomly removing $p \leq 300$ subjects from the original set and all their actions (that is up to 55% of the data): we observe that the optimum in the DB index remains around the same value $k = 5$. However, the SD is larger for all the leave-p-out cases, and for any given k or p, than for the analysis perform over the original data set. This variability gets larger the more data is randomly excluded. Of course, if too much of the data is removed ($p \geq 300$ subjects), the K-means algorithm is no longer able to retrieve the original optimum cluster structure, as can be inferred from the gradual disappearance of the local minimum in fig. S17 as p increases. We remind the reader that a data set lacking any cluster structure would render a monotonically decreasing DB index as a function of the number of clusters.

S.4.8. DB index

The Davies-Bouldin index, or DB index (30), is a metric for evaluating and comparing clustering algorithms. It is minimized by the optimum clustering scheme, that is to say, by the partition in a number of clusters such that it presents the minimum dispersion within each cluster, and the maximum distance between all pairs of clusters. In particular, this metric performs an internal evaluation, that is, the validation of the goodness of the clustering partition is made using quantities inherent to the data set. Hence, it does not do a validation against the 'ground truth'. We picked this particular validation method because in this context there isn't a known ground truth for types of players (or 'phenotypes').

Given a certain scheme or partition in N clusters, let C_i be a cluster of vectors, and let

$\vec{X}_\ell$ be an n-dimensional feature vector that represents subject ℓ (in our particular case, $n = 4$ dimensions), who is assigned to cluster C_i. The dispersion S_i within cluster C_i is calculated as

$$S_i = \frac{1}{T_i} \sum_{\ell=1}^{T_i} \|\vec{X}_\ell - \vec{A}_i\| \tag{1}$$

where $\vec{A}_i$ is the centroid of cluster C_i, $\|\vec{X}_\ell - \vec{A}_i\|$ denotes the Euclidean distance between the vector $\vec{X}_\ell$ and the centroid $\vec{A}_i$, and T_i is the size of cluster C_i (that is, the number of subjects assigned to that cluster).

Then for each pair of clusters i and j, we define the matrix

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{2}$$

where $M_{ij} = \|\vec{A}_i - \vec{A}_j\|$ is the separation between clusters i and j (that is, the distance between their corresponding centroids).

Thus, we can define the DB index as

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{3}$$

where $D_i = \max_{i \neq j} R_{ij}$.

S.4.9. Normalized Mutual Information Score

In order to compare the consistency between two independent runs of the K-means algorithm in terms of the individuals' composition of the clusters obtained, we use the Normalized Mutual Information Score (37), as implemented in the Python package SciKit Learn).

The Mutual Information is a measure of the similarity between two clustering (or labeling) systems U and V of the same data into disjoint subsets, and it is given by the relative entropy between the joint distribution and the product distribution. Mutual Information between clustering systems U and V is then defined as

$$MI(UV) = \sum_{i=1}^{U} \sum_{j=1}^{V} P(i,j) log \frac{P(i,j)}{P(i)P'(j)} \tag{4}$$

where $P(i)$ is the probability of a random sample occurring in cluster U_i and $P'(j)$ is the probability of a random sample occurring in cluster V_j.

To obtain a Normalized Mutual Information Score in such a way that it is bounded between 0 (no mutual information) and 1 (perfect correlation), the Mutual Information is normalized by $\sqrt{H(U) * H(V)}$, being $H(U)$ the entropy of the clustering system U, and $H(V)$ that of clustering V.

Note that this metric is independent of the absolute values of the labels: a permutation of the class or cluster label values will not change the score value in any way, and furthermore, it is symmetric, since switching the labels from clustering system U to clustering system V will return the same score value.

In fig. S18 we present the average value of Normalized Mutual Information, $\langle MI \rangle$, for any number of clusters in the original data, and in the randomized version of the data, over $2,000$ runs of the algorithm. We proceed as follows: we perform $(1,000)$ pair-wise comparison of clustering schemes obtain in different runs, calculating its corresponding score in each case, so then we can obtain an average. We observe how the score is significantly higher in the case of the actual data, than when comparing with the results from a randomized version of the data (for example, $\langle MI \rangle = 0.97$, $SD : 0.03$, vs $\langle MI \rangle = 0.59$, $SD : 0.18$ for actual and randomized results at $k = 5$ clusters), which indicates that the individuals composition of the clusters in any two runs of the algorithm on the real data are extremely correlated, but it is not the case for two runs over randomized data. Finally, we also report that the score is at its highest value for $k = 5$ clusters.

S.4.10. Age and gender by phenotype

The average (SD) age by phenotype is: for the Envious is $29.9(13.9)$; $32.5(13.7)$ for the Optimist; $32.0(16.8)$ for the Undefined; $32.29(14.1)$ for the Cooperators, and $30.7(13.8)$ for the Pessimist. We do not observe significant differences on the average age among different phenotypes nor with respect with the population average (31.3, SD: 14.3).

In fig. S19 we present the age distributions by phenotype, as they compare to the distribution for the whole population. We do not observe significant differences for any of the distributions by phenotype when comparing with that for the whole population nor by doing pair-wise comparisons of different phenotypes. The corresponding p-values for the KS-test (used to compare the probability distribution of two samples) of each possible pairwise combinations are non-significant: Envious vs Optimist: 0.31; Envious vs Undefined: 0.29; Envious vs Trustful: 0.32; Envious vs Pessimist: 0.81; Optimist vs Undefined: 0.57; Optimist vs Trustful: 0.99; Optimist vs Pessimist: 0.64; Undefined vs Trustful: 0.71; Undefined vs Trustful: 0.71; Undefined vs Pessimist:0.68; Trustful vs Pessimist: 0.67. Similarly, the p-values for all comparisons between clusters and the whole population are non-significant: Optimist vs all: 0.88; Envious vs all: 0.79; Undefined vs all: 0.68; Trustful vs all: 0.88; Pessimist vs all: 0.81.

The percentage of males for each phenotype is: 67% among the Envious, 64% among the Optimist, 64% among the Undefined, 61% among the Cooperators and 64% among the Pessimists (while the percentage of males for the whole populations is 64%). The z-scores of the comparison of gender distributions of each cluster vs the whole population by bootstrapping are all non-significant: Envious vs all: -0.036; Optimist vs all: -0.460; Undefined vs all: -0.260 ; Trustful vs all: -0.646; Pessimist vs all: -0.132.

S.4.11. Differences between experimental and numerical behavioral heatmaps

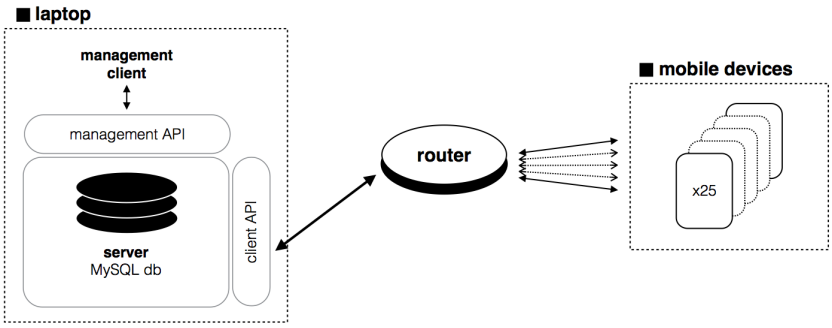
Assuming that each subject in our study plays using one and only one of the behavioural rules or phenotypes, and preserving the relative fractions of each one of them present in the population as found by the clustering algorithm, we can compute the differences between experimental and numerical (or inferred) behavioral heatmaps for each phenotype. In fig. S20 it can be seen that, even if occasionally the difference can reach up to 4 SD units for a particular (T, S) point, there is no systematic bias in any of the different heatmaps. The average difference in the aggregate case is of 1.39 SD units, while the difference by phenotype are: 1.91 SD units for Envious, 1.85 SD units for Optimist, 2.14 SD units for Pessimist, 1.79 SD units for Trustful, 1.12 SD units for Undefined. Thus none of the phenotypes presents an average difference beyond the 99% Confidence Interval (2.575 SD units). Indeed, only Pessimists present an average difference out of 95% Confidence Interval (1.96 SD units), the rest are below such standard threshold. We thus clearly show that the aggregation of the behavior of our volunteers into the proposed phenotypes is not significantly different from what we have obtained in the experiment.

S.4.12. Dependence of cooperation on $S - T$

Inspired by the population-level observation about the patterns described by lines parallel to $S = T$, and the fact that the population as a whole does not seem to distinguish between SH and SG, we studied cooperation as a function of the combined variable $(S - T)$. The results, represented in fig. S21, show a remarkable collapse of all curves into a single one, indicating that the aggregate cooperation level can be described by $(S - T)$, as previously pointed out by Rapaport (11, 13). In this respect, it is worth noting that $(S - T)$ represents the maximum possible payoff difference for any game. For very negative values of $(S - T)$, which corresponds to the PD game, the levels of cooperation are low but not zero, while for positive values (corresponding to HG) they are high, with intermediate, increasing values of

cooperation for the region $(S-T) \in [-10, 0]$, which roughly corresponds to a combination of the coordination and the anti-coordination games. This suggests that competition, in the sense of ending up being better off than one's counterpart, may be important for our experimental subjects.

Further, we check whether these results are reproduced from our interpretation of the clustering results and the corresponding simulations. In fig. S22 we plot together the results obtained from numerical simulations that use the experimentally obtained classification. As shown, by simply using the right fraction of each phenotype (behavioral rules) in the population, we can recover the observed diagonal symmetry, thus further confirming our 5-phenotype hypothesis.



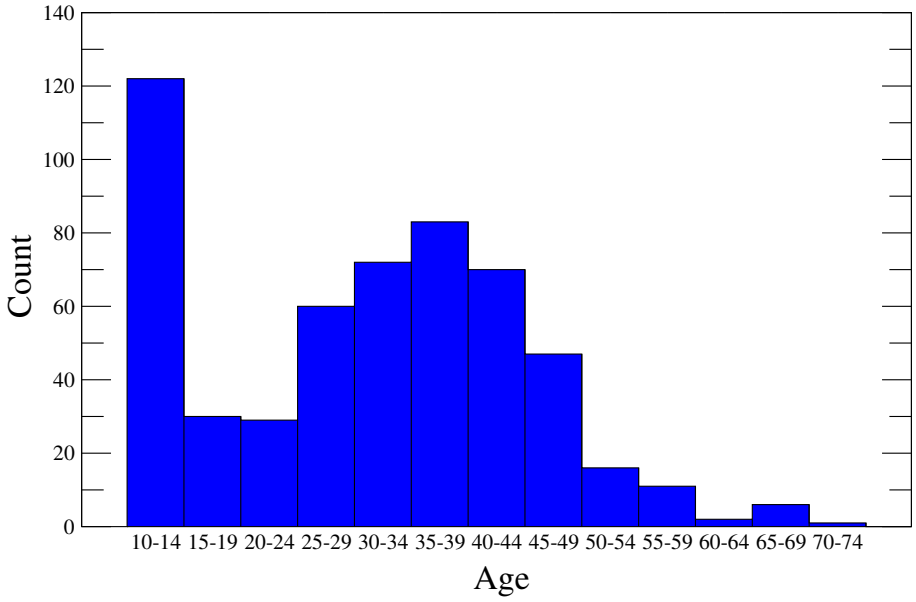


fig. S2. Age distribution of the participants in our experiment.

fig. S3. Screenshots of the tutorial shown to participants before starting the experiment, and feedback screen after a typical round of the game. See text for translation.

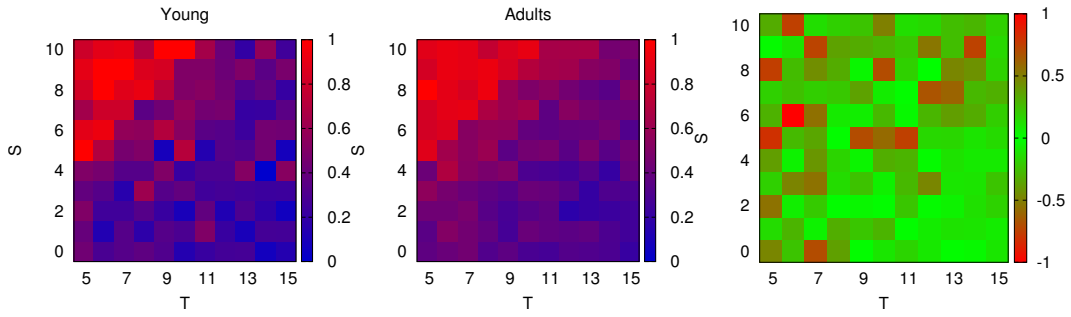


fig. S4. Fraction of cooperative actions for young ($\leq$ 15 years old) and adult players ($>$ 16 years old), and relative difference between the two heatmaps: (young-adults)/adults.

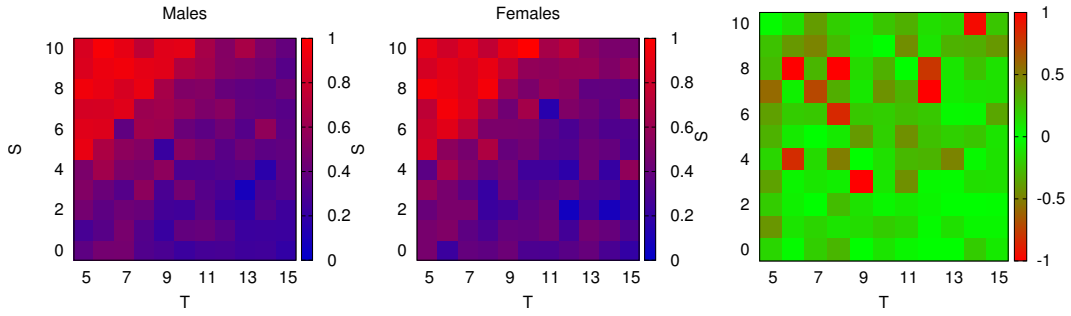


fig. S5. Fraction of cooperative actions for males and females separately, and relative difference between the two heatmaps: (males-females)/females.

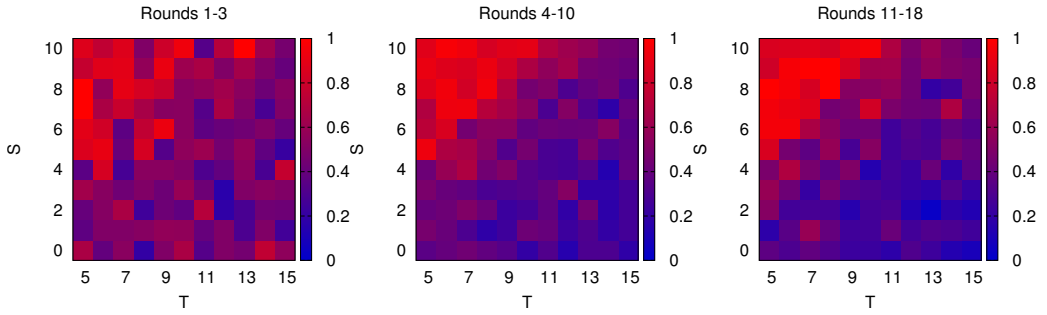


fig. S6. Fraction of cooperative actions separating by round number: for the first 1 to 3 rounds, 4 to 10 and last 11 to 18 rounds.

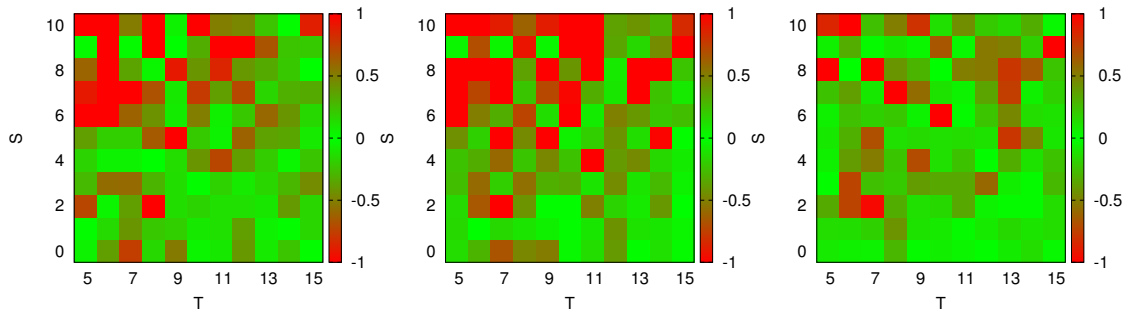


fig. S7. Relative difference in the fraction of cooperation heatmaps between groups of rounds. Left: (rounds 1 to 3 - rounds 4 to 10)/(rounds 4 to 10); Center: (rounds 1 to 3 - rounds 11 to 18)/(rounds 11 to 18) ; Right: (rounds 4 to 10 - rounds 11 to 18)/(rounds 11 to 18).

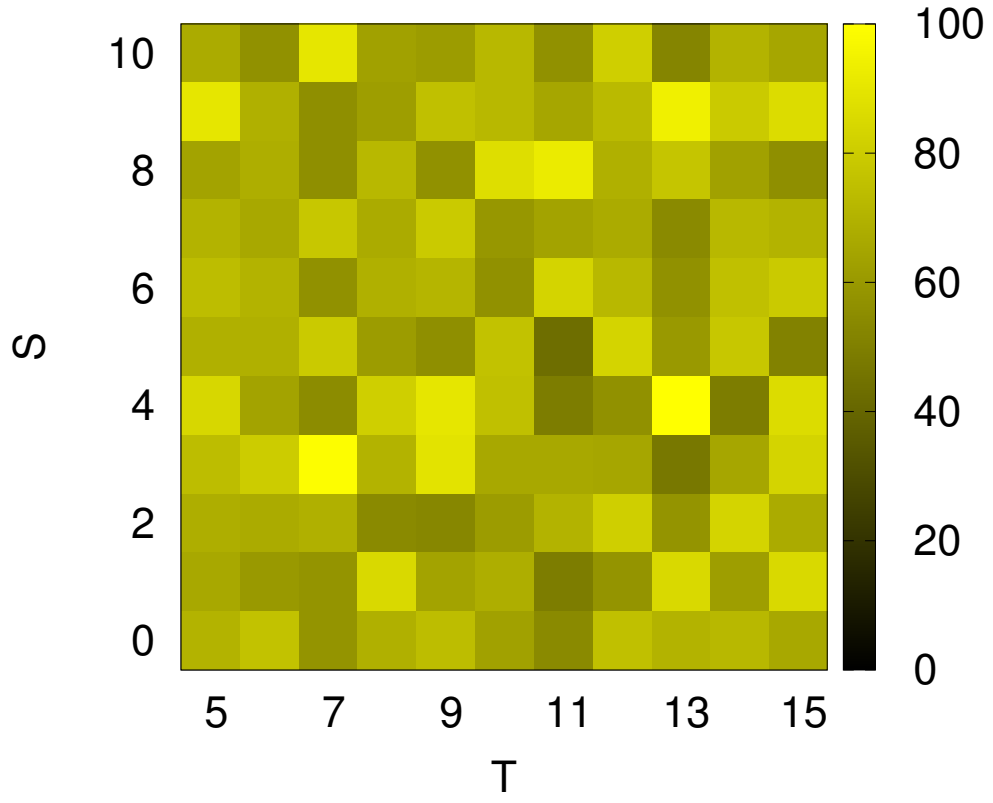


fig. S8. Total number of actions in each point of the (T, S)-plane, for all 541 participants in the experiment (the total number of game actions in the experiment adds up to $8,366$).

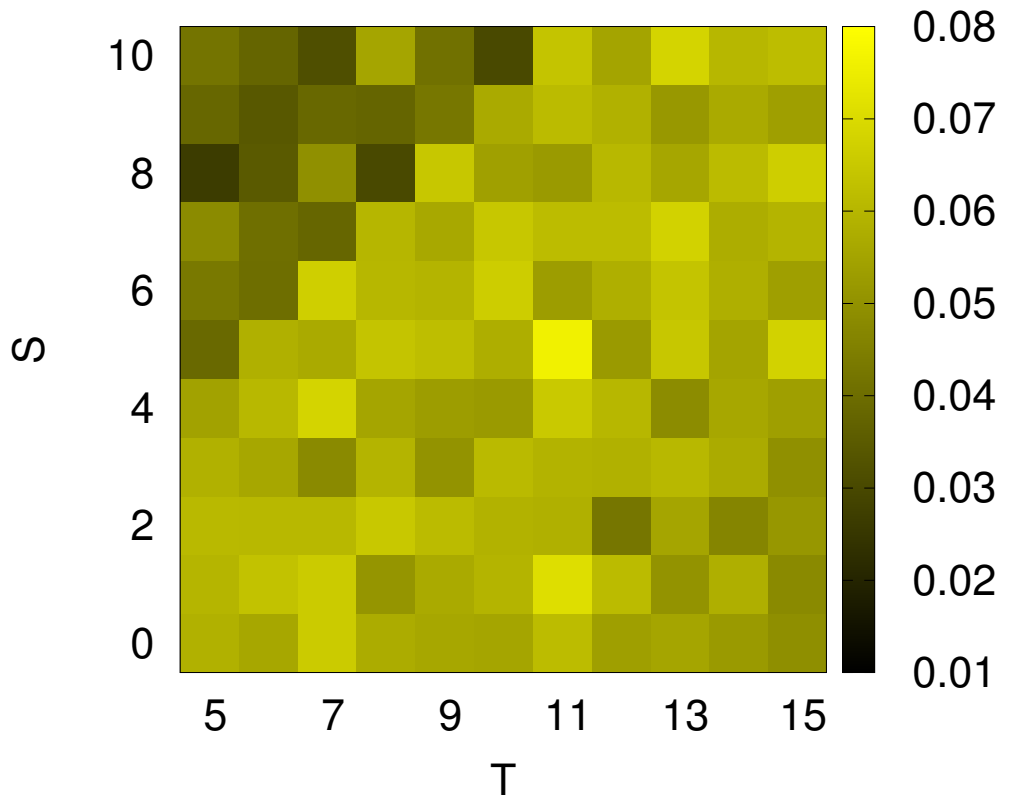


fig. S9. Standard Error of the Mean (SEM) fraction of cooperative actions in each point of the (T, S)-plane, for all the participants in the experiment. The HG regions leads to lower SEM of cooperation and that was expected given that two important types of phenotypes predict cooperation. To get a 95% Confidence Interval errors bars should be multiplied by 1.96.

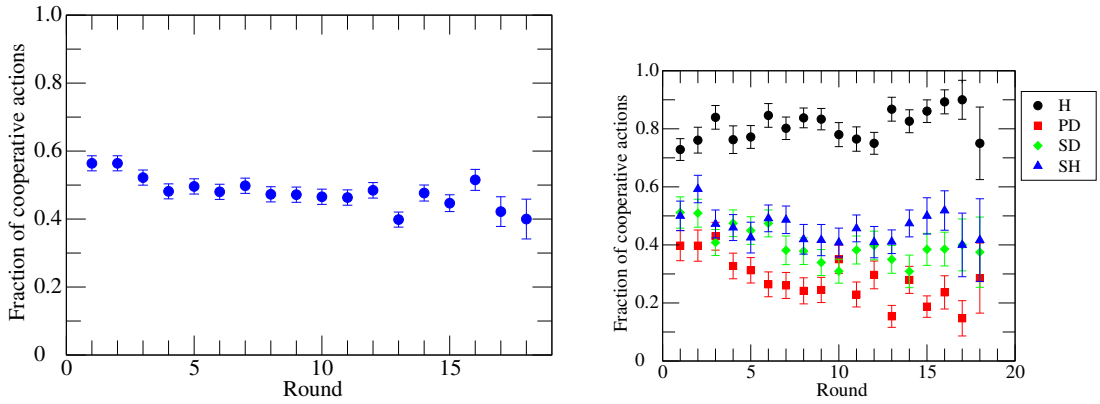


fig. S10. Average fraction of cooperative actions (and Standard Error of the Mean) among the population as a function of the round number overall (left) and separating the actions by game (right).

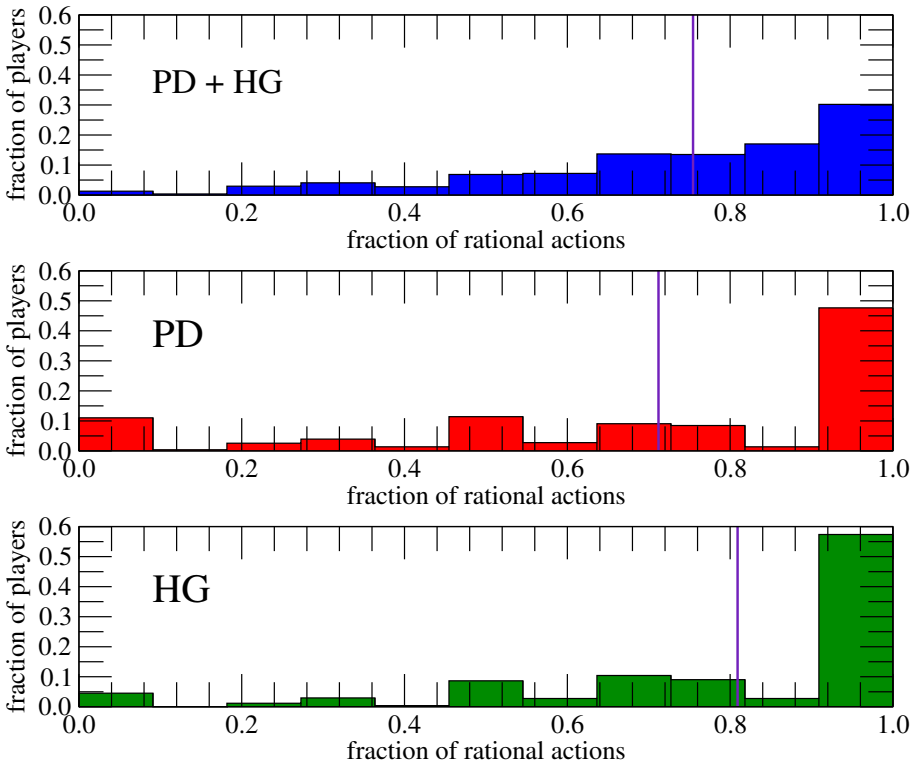


fig. S11. Distribution of fraction of rational actions among the 541 subjects of our experiment, when considering only their actions in the Harmony game (HG), or the Prisoner's Dilemma (PD), or both together. The purple line indicates the mean value.

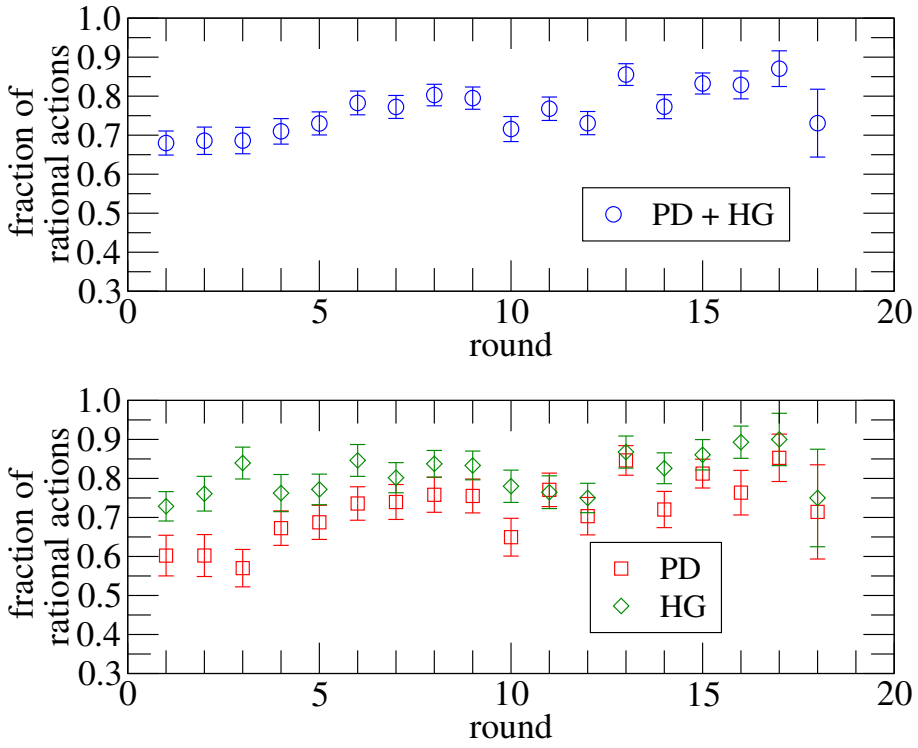


fig. S12. Fraction of rational actions as a function of the round number for the 541 subjects, defined by their actions in the Prisoner's Dilemma game (PD) and Harmony game (HG) together (top), and independently (bottom). The bars correspond to the Standard Error of the Mean.

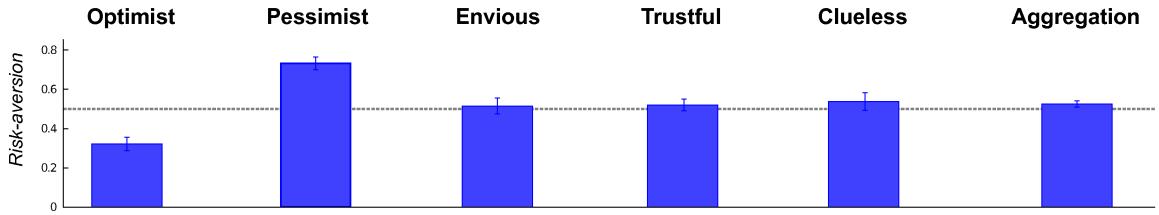


fig. S13. Values of risk-aversion averaged over the subjects in each phenotype. The phenotypes of Optimist and Pessimist show significantly lower and higher values than random expectation, respectively. Error bars indicate 95% Confidence Intervals.

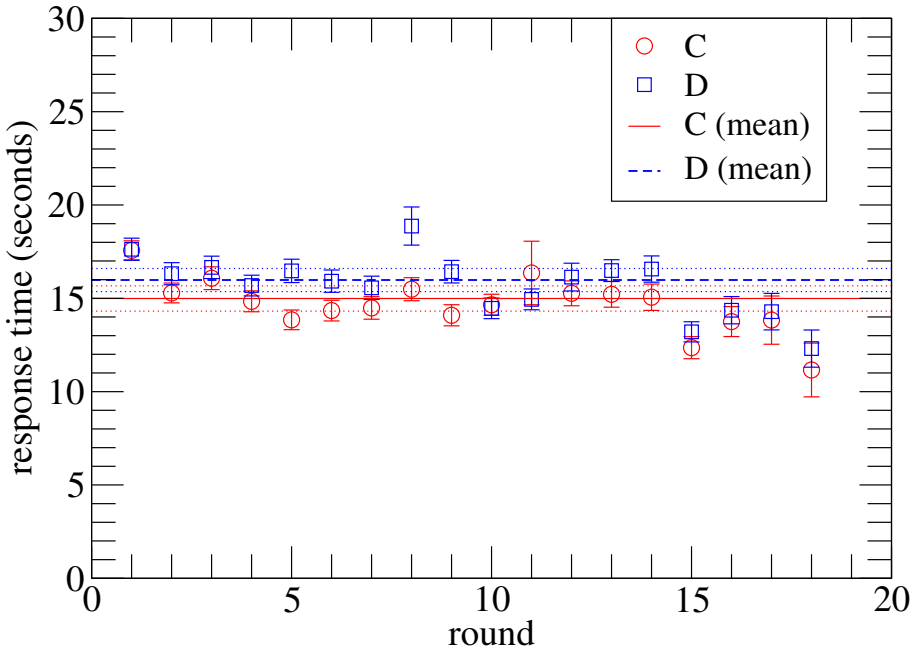


fig. S14. Average response times (and Standard Error of the Mean) as a function of the round number, for all the participants in the experiment, and separating the actions into cooperation (C) or defection (D).

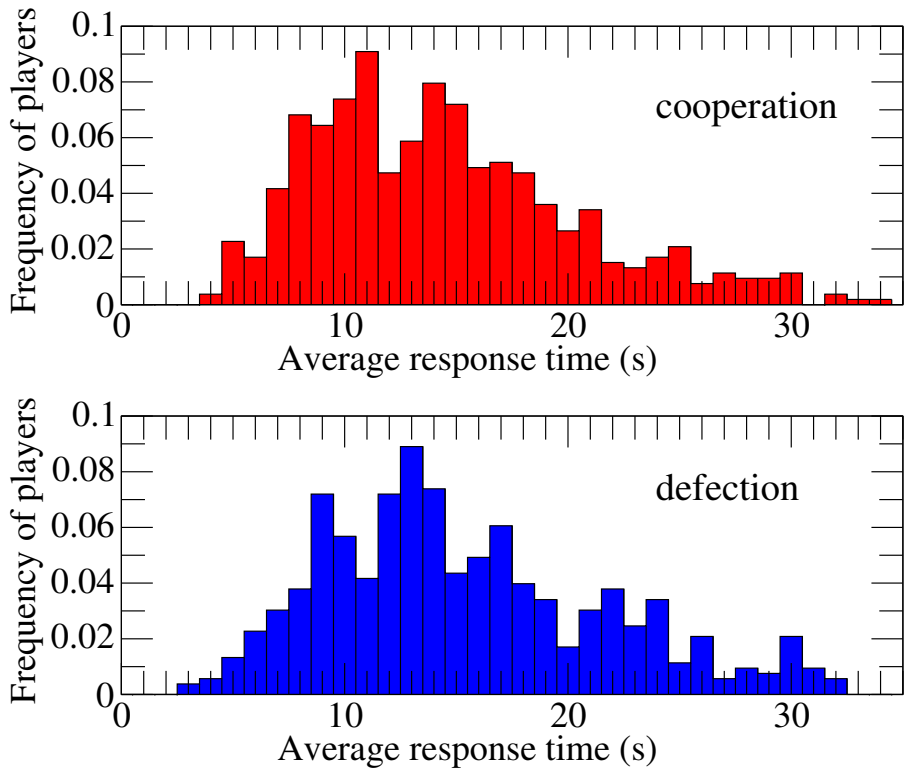


fig. S15. Distributions of response times for all the participants in the experiment, and separating the actions into cooperation (top) and defection (bottom).

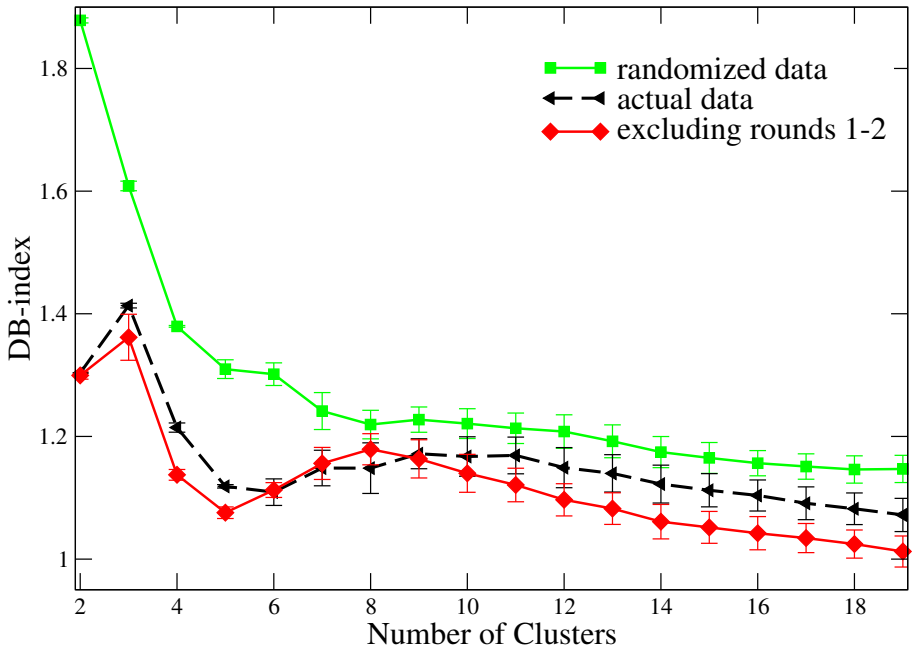


fig. S16. Testing the robustness of the results from the K-means algorithm. We present the average value of the DB-index over 200 independent runs of the algorithm on the data, as a function of the number of clusters (black). The optimum number of clusters is 5 (we note that, although a 6-cluster partition is also comparably good, the Standard Deviation (SD) is larger in that case, indicating less stability across different runs). We also show the results for the case of a randomization of the data (green). In this case, we observe that there is no local minimum, indicating a lack of cluster structure. Finally, we observe that when excluding the first two choices of every subject in our experiment (to account for excessive noise due to lack of experience), the position of the optimum is located in a clearer way at 5.

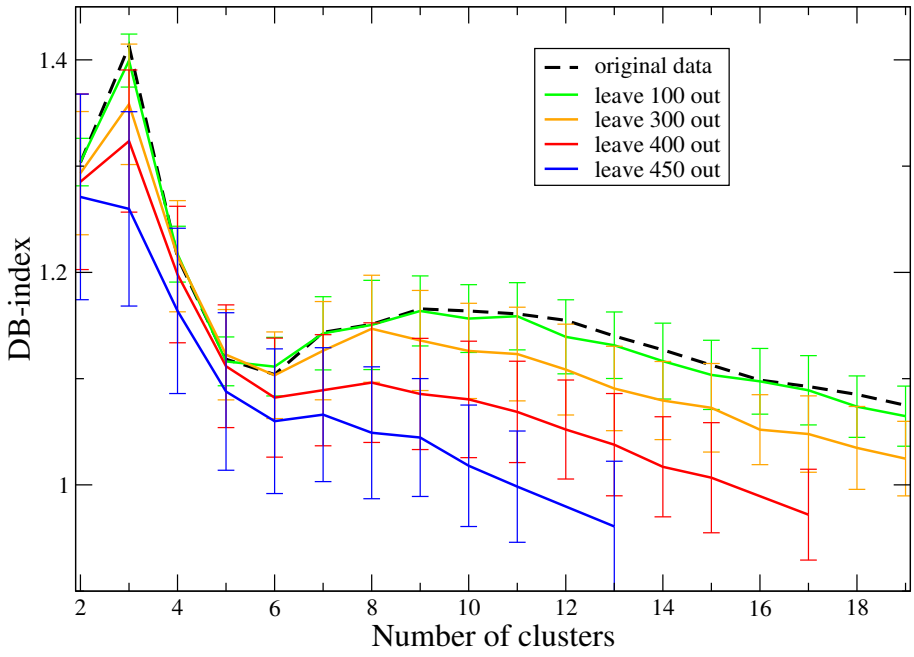


fig. S17. DB-index as a function of the number of clusters in the partition of our data (dashed black) as it compares to the equivalent results for different leave-p-out analyses. See text for details. The bars correspond to the Standard Deviation over the 200 independent realizations of the algorithm.

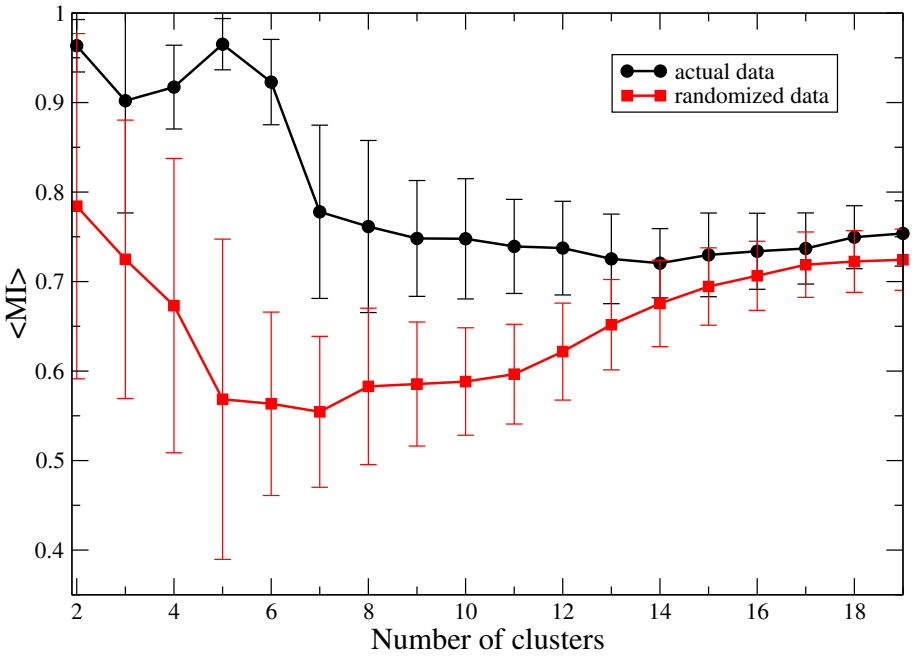


fig. S18. Average value for the Normalized Mutual Information Score, when doing pair-wise comparisons of the clustering schemes from $2,000$ independent runs of the K-means algorithm, both on the actual data, and on the randomized version of the data.

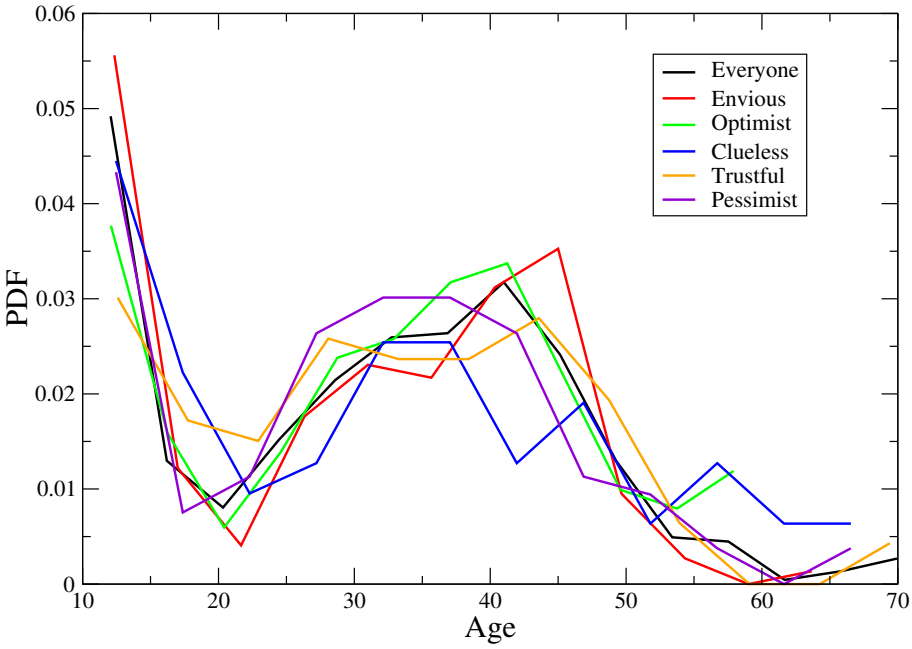


fig. S19. Age distribution for the different phenotypes, as it compares to the distribution of the whole population (black).

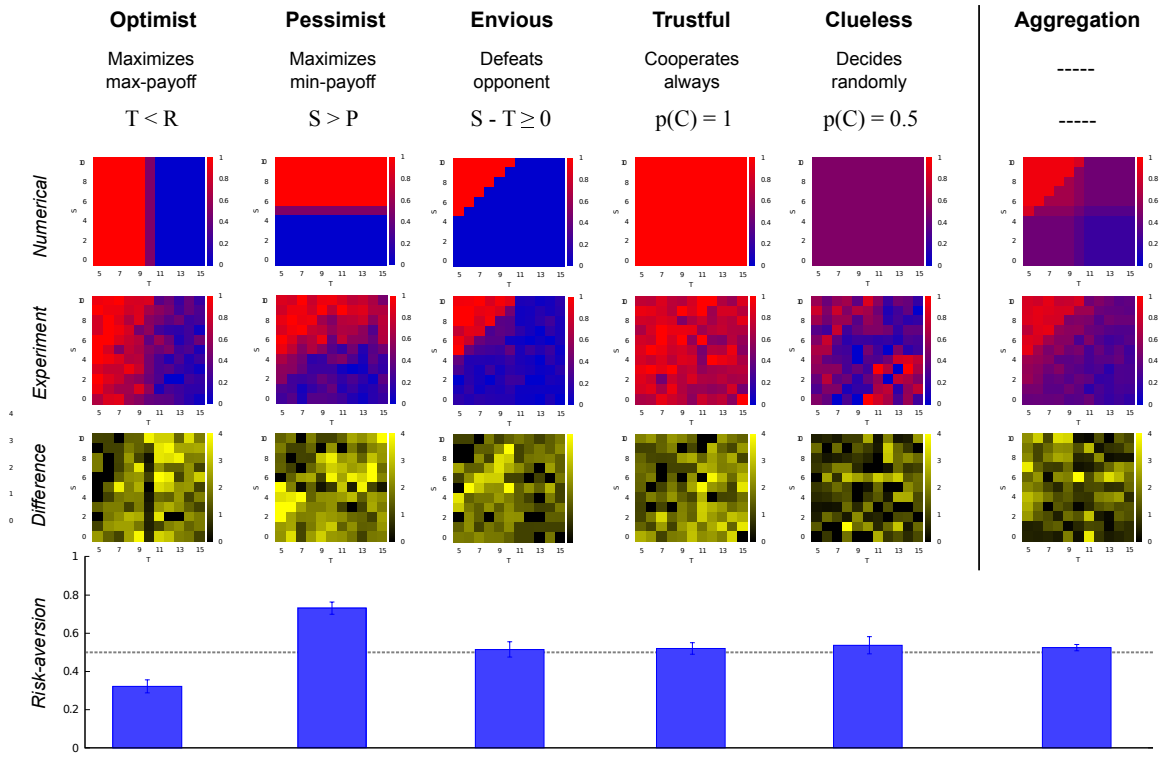


fig. S20. Difference between the experimental (second row) and numerical (or inferred, first row) behavioral heatmaps for each one of the phenotypes found by the K-means clustering algorithm, in units of SD. The difference between theory and experiment averaged over all (T, S)-plane is 1.91 SD units for Envious, 1.85 SD units for Optimist, 2.14 SD units for Pessimist, 1.79 SD units for Trustful, 1.12 SD units for Undefined and 1.39 SD units for the overall results in the Aggregation column.

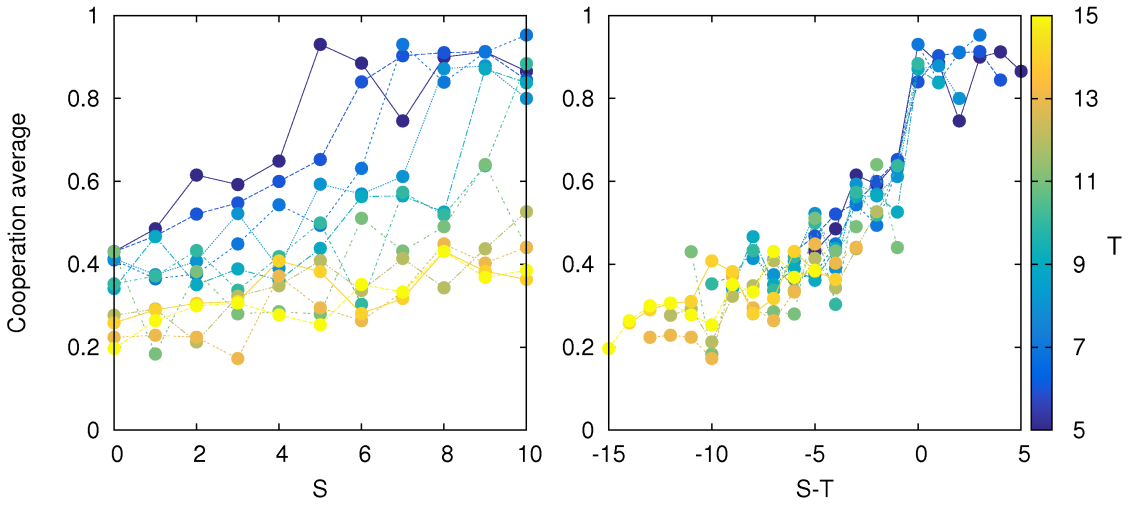


fig. S21. Average level of cooperation over all game actions and for different values of T (in different colours). We observe disparate results when cooperation fraction is represented as a function of S (left) but we find a nice collapse of all curves when cooperation level is expressed as a function $(S - T)$ (right).

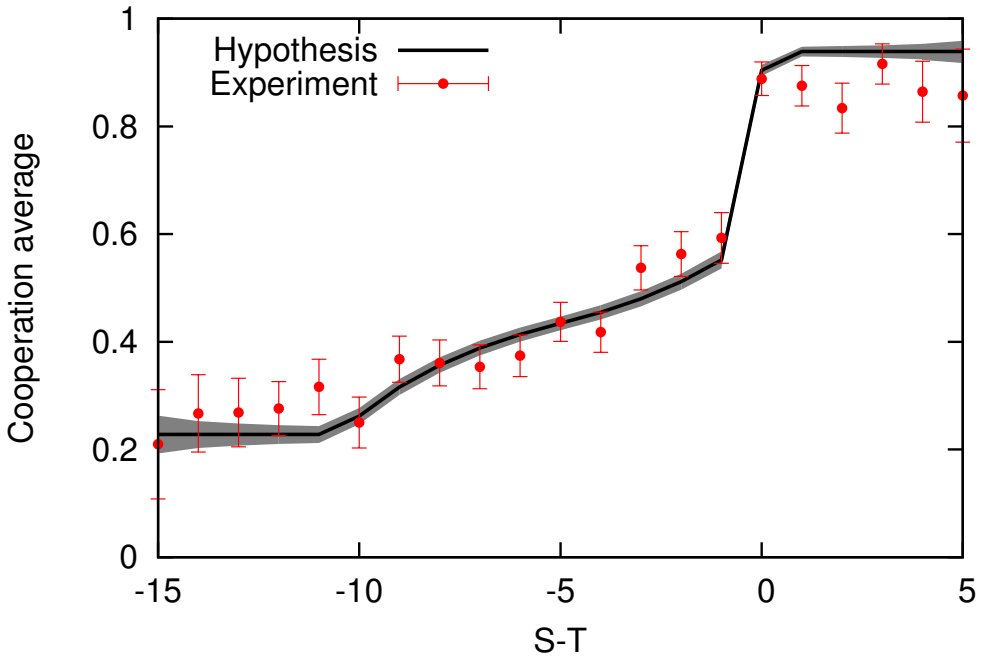


fig. S22. Average level of cooperation as a function of $(S-T)$ for both hypothesis and experiment. We consider the weight (number of decisions) in each cell when averaging over cells with same $(S-T)$. The error bars and the grey area represents a 95% Confidence Interval for the experimental points and the recreated curve respectively.